Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.018414
Article

Tech Science Press

Binaural Speech Separation Algorithm Based on Deep Clustering

Lin Zhou[1,*], Kun Feng[1], Tianyi Wang[1], Yue Xu[1] and Jingang Shi[2]

[1]School of Information Science and Engineering, Southeast University, Nanjing, 210096, China
[2]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, FI-90014, Finland
*Corresponding Author: Lin Zhou. Email: Linzhou@seu.edu.cn

Received: 08 March 2021; Accepted: 09 April 2021

Abstract: Neutral network (NN) and clustering are the two commonly used methods for speech separation based on supervised learning. Recently, deep clustering methods have shown promising performance. In our study, considering that the spectrum of the sound source has time correlation, and the spatial position of the sound source has short-term stability, we combine the spectral and spatial features for deep clustering. In this work, the logarithmic amplitude spectrum (LPS) and the interaural phase difference (IPD) function of each time frequency (TF) unit for the binaural speech signal are extracted as feature. Then, these features of consecutive frames construct feature map, which are regarded as the input to the Bi-directional long short-term memory (BiLSTM). The feature maps are converted to the high-dimensional vectors through BiLSTM, which are used to classify the time-frequency units by K-means clustering. The clustering index are combined with mixed speech signal to reconstruct the target speech signal. The simulation results show that the proposed algorithm has a significant improvement in speech separation and speech quality, since the spectral and spatial information are all utilized for clustering. Also, the method is more generalized in untrained conditions compared with traditional NN method e.g., deep neural network (DNN) and convolutional neural networks (CNN) based method.

Keywords: Binaural speech separation; K-means clustering; BiLSTM

1 Introduction

As a fundamental algorithm in signal processing, speech separation can improve the performance of whole speech signal processing system such as speech recognition and speech interface systems, which has a wide range of applications [1–3].

Speech separation is to separate target speech from background interferences including noise, reverberation and interfering speech. Human auditory system can extract the target sound source in complex acoustic environment, which inspires the research on the perception mechanism of the human auditory system. The researchers propose the concept of computational auditory scene analysis (CASA) [4] which provides new research ideas for speech separation. CASA divides the perception process of human auditory sense into two stages: segmentation and grouping. In the segmentation stage, the speech

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

is decomposed into auditory segments, and each segment corresponds to a partial description of an acoustic event in the auditory scene. In the grouping stage, the auditory segments from the same sound source are recombined to form a perceptual structure of auditory stream. Therefore, related researches focus on the segmentation algorithm and grouping algorithm.

A recent approach treats segmentation algorithm as a supervised learning problem, which includes the following components: learning machines, training targets and acoustic features. Thus, related research are carried out in the above three aspects.

Rickard [5] proposes the Degenerate Unmixing Estimation Technique (DUET) algorithm, which extracts the Inter-aural Time Difference (ITD) and Interaural Level Difference (ILD) as spatial features. These spatial features are clustered to classify Time-Frequency (TF) units based on azimuth of sound source. Roman et al. [6] uses ITD and ILD to assign Ideal Binary Mask (IBM) to each TF unit to accomplish separation. However, since IBM are only binary, the separated speech is unnatural. Cho et al. [7] post-processes the spatial features extracted by DUET, which increases the accuracy of the model. Kim et al. [8] proposes a channel weighting algorithm on phase difference, which can maintain speech quality under low Signal Noise Ratio (SNR) and reverberation conditions. Fan et al. [9] combines monaural LPS features and the binaural features for spectrum mapping. The simulation results show that approach can significantly outperform IBM-based speech segregation in terms of speech quality and speech intelligibility for noisy and reverberant environments. Dadvar et al. [10] proposed a binaural speech separation algorithm based on a reliable spatial feature of smITD+smILD which is obtained by soft missing data masking of binaural cues. And DNN maps the combined spectral and spatial features to a newly defined ideal ratio mask (IRM). It is shown that the proposed system outperforms the baseline systems in the intelligibility and quality of separated speech signals in reverberant and noisy conditions.

Deep learning has been introduced into speech separation due to the excellent learning ability. Narayanan et al. [11] combines Deep Neural Networks (DNN) with speech separation for the first time. In this study, the binaural speech signals are filtered through the Gammatone filter to obtain the corresponding TF unit, and extracted ITD, ILD and GFCC as features for DNN training. Simulation results show that DNN, with non-linear mapping capability, has a good performance in speech separation. In order to solve the noise generalization, Xu et al. [12] introduces a noise perception training process on DNN, which achieves good performance even to the untrained noise. Xiao uses DNN to predict the logarithmic amplitude and cross-correlation parameters of clean speech, which improves the performance of de-reverberation. Zhang et al. [13] and Luo et al. [14] conduct study on Long Short-Term Memory (LSTM), which significantly improves performance of speaker generalization. Hershey proposes a Deep Clustering (DPCL) method for monaural music separation. But this algorithm is lack of the spatial information for the binaural signal. Zhao et al. [15] proposes a two-stage DNN structure, which is used for noise reduction and de-reverberation respectively. Venkatesan et al. [16] proposes a binaural speech separation and speaker recognition system based on Deep Recurrent Neural Network (DRNN), while DRNN has the problems of gradient disappearance and gradient explosion. Liu et al. [17] proposes a binaural speech separation algorithm based on iterative-DNN to retrieve two concurrent speech signals in a room environment. Objective evaluations in terms of PESQ and STOI showed consistent improvement over baseline methods. Zermini et al. [18] proposes a speech separation algorithm based on Convolutional Neural Networks (CNN), which greatly improves the signal to distortion ratio (SDR). Luo et al. [19] proposes a fully-convolutional time-domain audio separation network (Conv-TasNet), a deep learning framework for end-to-end time-domain speech separation. Aiming at the problems caused by the non-parallelism of LSTM and BiLSTM, Conv-TasNet replaces LSTM with TCN.

In previous studies, only the information of the current TF unit is utilized. Considering that the speech spectrum for the same voice source has time correlation, and the spatial position of the voice source has short-term stability, this paper utilizes BiLSTM as the encoder to map the logarithmic amplitude spectrum and the inter-aural phase difference (IPD) function to the high-dimensional vectors, which are used to classify the time-frequency units by K-Means clustering. This approach combines the spectral and spatial features of consecutive frames to perform speech separation. Compared with CNN-based networks, it has an improvement in SAR, SIR and SDR in trained and untrained environment.

The remainder of the paper is organized as follows. Section 2 presents an overview of binaural speech separation system based on deep clustering and the extraction of spectral feature and spatial feature. Section 3 describes the structure and training of Deep Clustering Networks. The simulation results and analysis are provided in Section 4. The conclusion is drawn in Section 5.

2 System Overview and Feature Extraction

In the training, the logarithmic amplitude spectrum and phase difference function of the binaural speech signal [20] are combined to train the encoding layer of the deep clustering network [21], and the deep clustering model maps the features of the binaural speech to high-dimension vectors. During the test, the features of the binaural speech are mapped to high-dimension vectors, and the time-frequency units are classified through clustering to obtain a binary mask matrix [22], thereby separating the mixed speech. The structure of the algorithm is shown in Fig. 1:

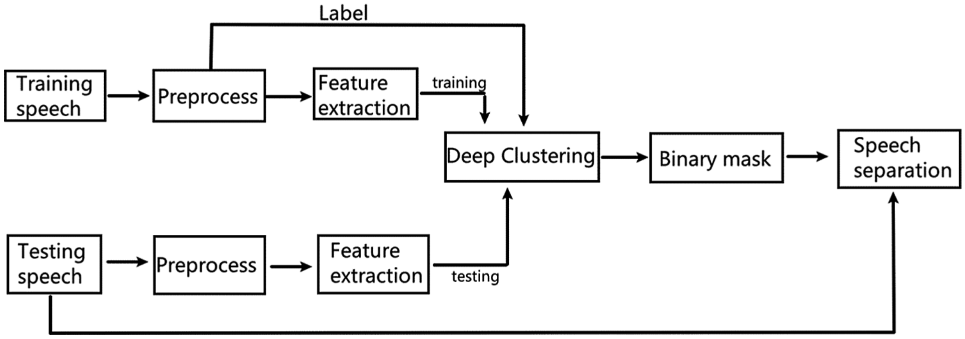

Figure 1: Schematic diagram of binaural speech separation based on deep clustering

2.1 Preprocess

The model for binaural speech signals in reverberant and noisy environments can be formulated as the Eqs. (1) and (2):

$$x_L(t) = h_{1,L} * s_1(t) + h_{2,L} * s_2(t) + n_L(t) \tag{1}$$

$$x_R(t) = h_{1,R} * s_1(t) + h_{2,R} * s_2(t) + n_R(t) \tag{2}$$

where $x_L(t)$ and $x_R(t)$ are defined as a pair of binaural signals; $s_1(t)$ and $s_2(t)$ represent two speech sources; h_L and h_R are the binaural room impulse response (BRIR) for left and right ear respectively for each speech source; $n_L(t)$ and $n_R(t)$ are additive noise for each ear, which are irrelevant to each other.

2.2 Feature Extraction

Short-time Fourier transform (STFT) [23] is conducted on the binaural speech signal after windowing.

$$X_L(\tau,\omega) = \sum_{n=0}^{N-1} x_L(\tau,n)e^{-j\omega n} \tag{3}$$

$$X_R(\tau,\omega) = \sum_{n=0}^{N-1} x_R(\tau,n)e^{-j\omega n} \tag{4}$$

where $X_L(\tau,\omega)$ and $X_R(\tau,\omega)$ represent the spectrum of binaural signal at time τ and frequency ω.

Logarithmic spectrum of $X_L(\tau,\omega)$, $\log_{10}| X_L(\tau,\omega)|$, is selected spectral feature.

Logarithmic spectrum only relies on the amplitude, which ignores the spatial information brought by the binaural signal, thus Inter-aural Phase Difference (IPD) is calculated as spatial feature.

IPD, $\varphi(\tau,\omega)$, is defined as the phase difference of $X_L(\tau,\omega)$ and $X_R(\tau,\omega)$, and can be described as:

$$\varphi(\tau,\omega) = \varphi_L(\tau,\omega) - \varphi_R(\tau,\omega) \tag{5}$$

where $\varphi_L(\tau,\omega)$, $\varphi_R(\tau,\omega)$ represents the phase of $X_L(\tau,\omega)$ and $X_R(\tau,\omega)$ respectively, which are defined as:

$$X_L(\tau,\omega) = |X_L(\tau,\omega)|e^{j\varphi_L(\tau,\omega)} \tag{6}$$

$$X_R(\tau,\omega) = |X_R(\tau,\omega)|e^{j\varphi_R(\tau,\omega)} \tag{7}$$

Cosine and sine function of IPD is formulated as:

$$\cos IPD(\tau,\omega) = \cos(\varphi(\tau,\omega)) \tag{8}$$

$$\sin IPD(\tau,\omega) = \sin(\varphi(\tau,\omega)) \tag{9}$$

The logarithmic spectrum and IPD function are constructed a new feature vector for each TF unit:

$$C(\tau) = [\,\log_{10}|X_L(\tau,\omega)|, \quad \cos IPD(\tau,\omega), \quad \sin IPD(\tau,\omega)\,] \tag{10}$$

The feature vectors of every T frames are combined to obtain the feature map shown in Eq. (11):

$$\boldsymbol{C} = [C(1), C(2), \ldots, C(T)] \tag{11}$$

Fig. 2 shows the logarithmic spectrum of the mixed speech signal and each target speech signal. The spectrum of the mixed speech is roughly the sum of the two targets from the outline. Fig. 3 is a schematic diagram of *sinIPD* and *cosIPD* of a mixed signal. It is obvious be seen that the distribution of *sinIPD* and *cosIPD* varies with frequency.

3 Binaural Speech Separation Based on Deep Clustering

Deep clustering is mainly composed of encoding layer and clustering layer. Only the encoding layer is used during training. In the testing, the high-dimension vectors are obtained through the encoding layer, and each TF unit is classified in clustering layer. The structure is shown in Fig. 4.

The encoding layer is composed of a BiLSTM [24,25], a dropout layer and a fully connected layer. The dimension of input feature map, $\boldsymbol{C} \in R^{3F\times T}$. Here, F represents the number of STFT points and T is the number of frames. The hidden layer of BiLSTM is set to 600, and the fully connected layer maps each time-frequency unit to the feature space of the K-dimension [26] embedding space.

In the training stage, the speech signal is preprocessed to obtain the logarithmic spectrum, *sinIPD* and *cosIPD* to construct the feature map, which is then sent to the BiLSTM. After the encoding layer, feature map is mapped to the K-dimension embedding space, as shown in Eq. (12):

$$V = f(C) \tag{12}$$

where $V \in R^{TF \times K}$ represents the TF unit matrix mapped to the high-dimension space through the encoding layer. Let the 2-norm of the row vector in V be 1, and VV^T represents the affinity matrix. Let $Y = \{y_{\tau*\omega,m}\}$ denote the belongs of each TF unit to each sound source, $Y \in R^{TF \times M}$ where M is the number of speakers. According to the masking based speech separation method, $y_{\tau*\omega,m} = 1$ when the amplitude of the m-th speaker is greater than other speakers, otherwise $y_{\tau*\omega,m} = 0$. YY^T is the affinity matrix of the classification results, with a value of 0 or 1. The loss function is expressed as Eq. (13):

$$J = \left\| VV^H - YY^H \right\|^2 \tag{13}$$

Eq. (13) makes the distance between the same speaker as small as possible, and the distance of different speakers larger. Then, speaker-independent speaker separation is achieved.

The BP algorithm [27] is used to implement the network training. In the testing, K-Means clustering is performed on high-dimension to realize the classification. The TF unit belongs to the same speaker are combined to obtain the separated target speech.

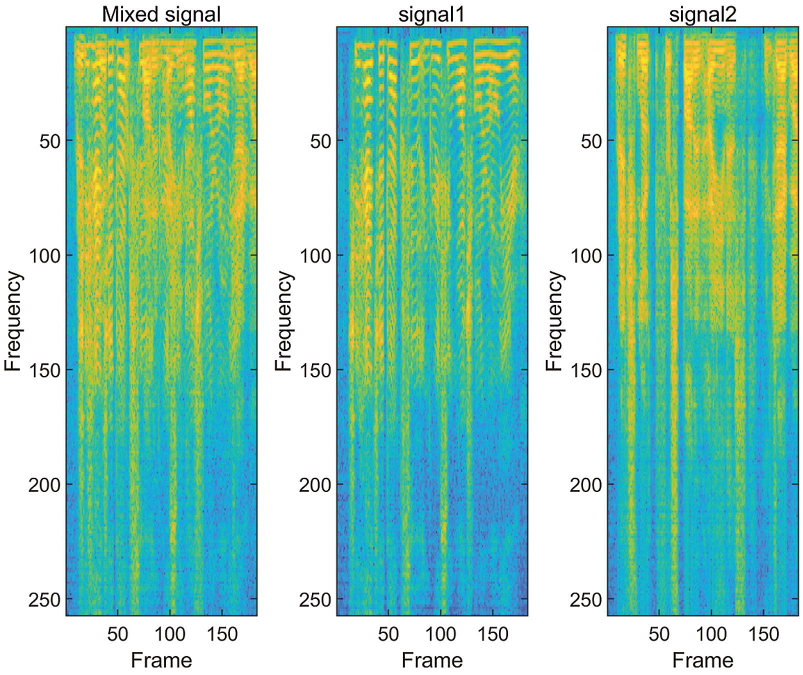

Figure 2: Speech signal logarithmic amplitude spectrum

4 Simulation and Result Analysis

4.1 Simulation Setup

HRIR is convolved with mono speech signals to obtain a directional binaural signal [28]. The mono speech signals are taken from the TIMIT database, which contains 630 speakers, each has 10 sentences. The sampling rate is 16 kHz. The azimuth varies between [–90°, 90°] with the step of 5°. The binaural speech signals with different azimuth are added to obtain the mixed speech for separation. The placement of two speech sources are depicted in Fig. 5.

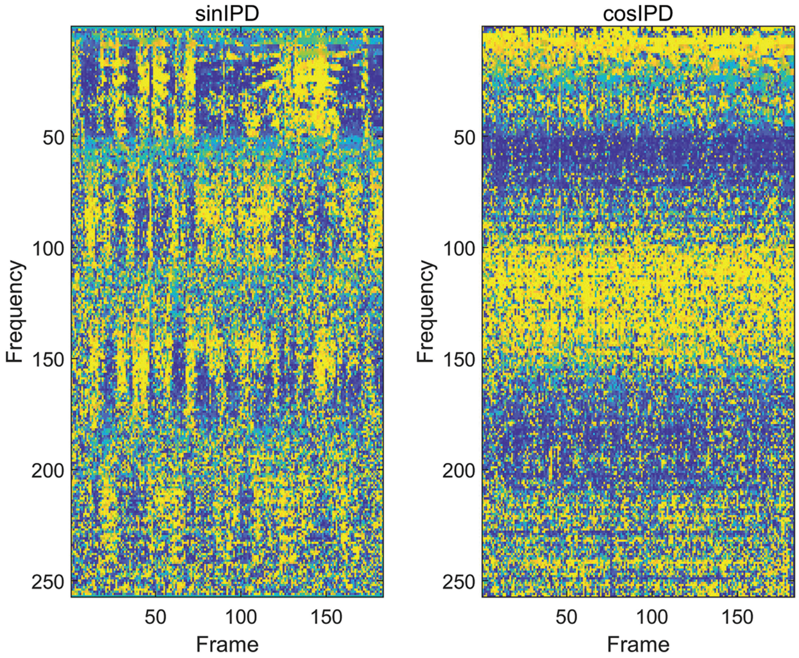

Figure 3: Schematic diagram of interaural phase difference of mixed speech signal

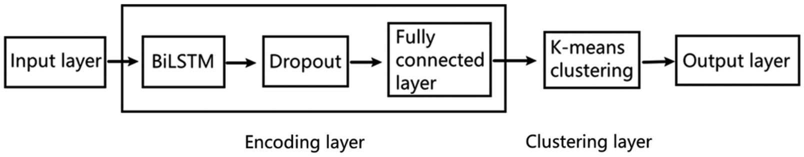

Figure 4: The architecture of deep clustering

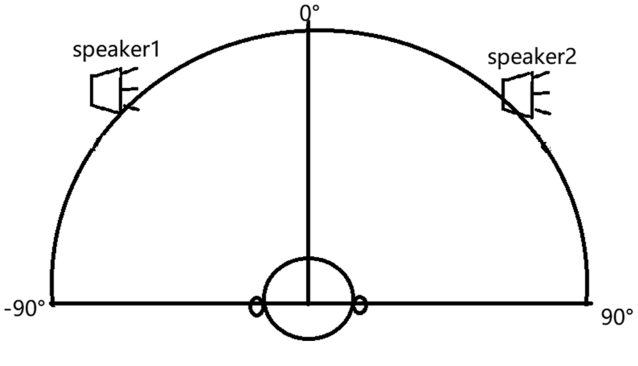

Figure 5: The spatial configuration of sources and receiver

Also, the Gaussian white noise is added to the binaural mixed speech as the ambient noise. There are 4 types of SNRs for training, which are 0 dB, 10 dB, 20 dB and no noise. For SNR generalization, the SNR for testing includes 5 dB and 15 dB. The total training sentence is about 80 hours. The testing speech is different from the training speech. Therefore, this simulation can be regarded as speaker-independent speech separation.

The reverberation time (RT60) for training are 0.2 s and 0.6 s. The RT60 for testing is 0.3 s, which verifies the generalization of the proposed algorithm to reverberation.

At the same time, in order to distinguish the noise segment from the silent segment during training, Voice Activity Detection (VAD) [29] is added and the TF unit less than the threshold will not be assigned to any speaker.

Sources to Artifacts Ratio (SAR), Source to Distortion Ratio (SDR), Source to Interferences Ratio (SIR) [30] and Perceptual Evaluation of Speech Quality (PESQ) are used to evaluate the performance of speech separation. SAR, SIR, SDR are mainly used to evaluate the performance of speech separation, PESQ is correlated with the speech quality, the value is in a range of 0 to 5.

We compare the performance of the proposed method, which is binaural speech separation based on deep clustering, with several other related methods for binaural speech separation. The two algorithms involved in the comparison are: DNN-based method with IBM, and CNN-based method.

4.2 Evaluation and Analysis

First, we evaluate the performance of the above algorithms in the noisy environment. SAR, SIR, SDR and PESQ of different algorithms are shown in Tabs. 1–4.

Table 1: SAR of different algorithms in noisy environment

SNR(dB)	IBM-DNN	CNN	Deep Clustering
0	0.07	2.02	1.57
5	2.71	4.54	4.02
10	6.02	6.95	7.15
15	7.81	8.01	8.54
20	8.34	8.77	9.12
Noiseless	8.85	9.03	9.44

Table 2: SIR of different algorithms in noisy environment

SNR(dB)	IBM-DNN	CNN	Deep Clustering
0	14.42	15.19	14.79
5	15.14	16.01	16.18
10	15.98	16.45	16.92
15	16.41	16.70	17.01
20	16.71	16.87	17.35
Noiseless	17.14	17.02	17.58

According to the performance comparison, in the low SNR, the performance of the proposed algorithm is close to that of CNN, while in the high SNR, deep clustering significantly improves the separation performance compare with the IBM-DNN and CNN. Also, for the unmatched SNR, 5 dB and 15 dB, the proposed algorithm maintains the speech separation and speech quality.

Table 3: SDR of different algorithms in noisy environment

SNR(dB)	IBM-DNN	CNN	Deep Clustering
0	−0.77	1.54	0.79
5	3.02	4.41	4.16
10	5.31	6.02	7.14
15	6.95	7.21	8.15
20	7.52	7.85	9.02
Noiseless	7.96	8.31	9.79

Table 4: PESQ of different algorithms in noisy environment

SNR(dB)	IBM-DNN	CNN	Deep Clustering
0	1.42	1.85	1.67
5	1.7	2.07	1.94
10	1.79	2.17	2.11
15	1.95	2.24	2.25
20	2.21	2.45	2.39
Noiseless	2.41	2.57	2.52

At the same time, we analyze the reverberation generalization of the proposed method and the CNN method. The RT60 of testing data is 0.3 s, which differs from that of training data. The comparison results are shown in Tabs. 5–7.

Table 5: SAR of two methods in reverberation environment (RT60 is 0.3 s)

SNR(dB)	CNN	Deep Clustering
0	1.89	1.32
5	4.07	3.95
10	6.61	6.70
15	7.45	7.79
20	8.26	8.71

The separation performance of the proposed algorithm is much better than that of the CNN method under non-matching reverberation, indicating that the reverberation generalization of the separation method based on deep clustering.

Table 6: SIR of two methods in reverberation environment (RT60 is 0.3 s)

SNR(dB)	CNN	Deep Clustering
0	14.77	14.51
5	15.82	15.94
10	15.91	16.41
15	16.54	16.63
20	16.68	16.72

Table 7: SDR of two methods in reverberation environment (RT60 is 0.3 s)

SNR(dB)	CNN	Deep Clustering
0	1.02	0.34
5	3.57	3.46
10	5.21	6.71
15	6.57	7.35
20	7.25	8.07

5 Conclusion

In this paper, we presented speech separation algorithm based on deep clustering. Considering that the frequency spectrum of the speech signal has time correlation, and the spatial position of the speech signal has short-term stability, the proposed algorithm combines the spectral and spatial features to form the feature map, which are sent to BiLSTM. After the encoding layer, feature map is mapped to high-dimensional vectors, which are used to classify the TF units by K-Means clustering. Speech separation based on deep clustering has shown its ability to improve speech separation and speech quality. At the same time, the proposed algorithm also maintains performance in unmatched SNR and reverberant environment, demonstrating noise and reverberation generalization.

Funding Statement: This work is supported by the National Key Research and Development Program of China under Grant 2020YFC2004002 and Grant 2020YFC2004003, and the National Nature Science Foundation of China (NSFC) under Grant No. 61571106.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Bao, Q. Shao, X. Zhang, J. Jiang, Y. Xie *et al.,* “A novel system for recognizing recording devices from recorded speech signals,” *Computers, Materials & Continua*, vol. 65, no. 3, pp. 2557–2570, 2020.

[2] J. Jo, H. Kim, I. Park, C. B. Jung and H. Yoo, “Modified viterbi scoring for HMM-based speech recognition,” *Intelligent Automation & Soft Computing*, vol. 25, no. 2, pp. 351–358, 2019.

[3] J. Park and S. Kim, “Noise cancellation based on voice activity detection using spectral variation for speech recognition in smart home devices,” *Intelligent Automation & Soft Computing*, vol. 26, no. 1, pp. 149–159, 2020.

[4] A. S. Bregman, *Auditory scene analysis the perceptual organization of sound*. Cambridge, MA: The MIT Press, 1994.

[5] S. Rickard, *The DUET blind source separation algorithm*. Dordrecht, Netherlands: Springer, 2007.

[6] N. Roman, D. L. Wang and G. J. Brown, “Speech segregation based on sound localization,” *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[7] N. Cho and C. J. Kuo, “Enhanced speech separation in room acoustic environments with selected binaural cues,” *IEEE Transactions on Consumer Electronics*, vol. 55, no. 4, pp. 2163–2171, 2009.

[8] C. Kim, K. Kumar and B. Raj, “Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain,” in *INTERSPEECH-2009*. Brighton, United Kingdom, pp. 2495–2498, 2009.

[9] N. Fan, J. Du and L. Dai, “A regression approach to binaural speech segregation via deep neural network,” in *2016 10th Int. Sym. on Chinese Spoken Language Processing (ISCSLP)*, Tianjin, China, pp. 1–5, 2016.

[10] P. Dadvar and M. Geravanchizadeh, “Robust binaural speech separation in adverse conditions based on deep neural network with modified spatial features and training target,” *Speech Communication*, vol. 108, no. 12, pp. 41–52, 2019.

[11] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, pp. 7092–7096, 2013.

[12] Y. Xu, J. Du and L. R. Dai, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[13] X. L. Zhang and D. L. Wang, “Boosting contextual information for deep neural network based voice activity detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2015.

[14] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux and N. Mesgarani, “Deep clustering and conventional networks for music separation: Stronger together,” in *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, pp. 61–65, 2017.

[15] Y. Zhao, Z. Wang and D. Wang, “A two-stage algorithm for noisy and reverberant speech enhancement,” in *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 5580–5584, 2017.

[16] R. Venkatesan and A. B. Ganesh, “Binaural classification-based speech segregation and robust speaker recognition system,” *Circuits Systems & Signal Processing*, vol. 37, no. 9, pp. 1–29, 2017.

[17] Q. Liu, Y. Xu, P. Jackson, W. Wang and P. Coleman, “Iterative deep neural networks for speaker-independent binaural blind speech separation,” in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 541–545, 2018.

[18] A. Zermini, Q. Kong and Y. Xu, “Improving reverberant speech separation with binaural cues using temporal context and convolutional neural networks,” in *Int. Conf. on Latent Variable Analysis and Signal Separation*, Cham, Switzerland: Springer, pp. 361–371, 2018.

[19] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[20] R. Nishimura, “Information hiding into interaural phase differences for stereo audio signals,” in *2009 Fifth Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, Kyoto, Japan, pp. 1189–1192, 2009.

[21] J. R. Hershey, Z. Chen, J. Le Roux and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 31–35, 2016.

[22] I. Misssaoui and Z. Lachiri, “Speech separation using DUET and binary masking with temporal smoothing in cepstral domain,” in *2013 World Congress on Computer and Information Technology (WCCIT)*, Sousse, Tunisia, pp. 1–4, 2013.

[23] A. Wang, G. Bi and B. Li, “Blind separation method of overlapped speech mixtures in STFT domain with noise and residual crosstalk suppression,” in *12th IEEE Int. Conf. on Control and Automation (ICCA)*, Kathmandu, Nepal, pp. 876–880, 2016.

[24] L. Zhou, S. Lu, Q. Zhong, Y. Chen and Y. Tang, “Binaural speech separation algorithm based on long- and short-time memory networks,” *Computers Materials & Continua*, vol. 63, no. 3, pp. 1373–1386, 2020.

[25] L. Ding, L. Li, J. Han, Y. Fan and D. Hu, “Detecting domain generation algorithms with bi-LSTM,” *Computers, Materials & Continua*, vol. 61, no. 3, pp. 1285–1304, 2019.

[26] Y. Xie, K. Xie, Z. Wu and S. Xie, “Underdetermined blind source separation of speech mixtures based on K-means clustering,” in *2019 Chinese Control Conf. (CCC)*, Guangzhou, China, pp. 42–46, 2019.

[27] Y. Li, Y. Fu, H. Li and S. Zhang, “The improved training algorithm of back propagation neural network with self-adaptive learning rate,” in *Int. Conf. on Computational Intelligence and Natural Computing*, Wuhan, China, pp. 73–76, 2009.

[28] MIT HRTF Database [DB/OL]. [Online]. Available: ftp://sound.media.mit.edu/pub/Data/KEMAR, 1996-06-20.

[29] Q. Liu and W. Wang, “Blind source separation and visual voice activity detection for target speech extraction,” in *2011 3rd Int. Conf. on Awareness Science and Technology (iCAST)*, Dalian, China, pp. 457–460, 2018.

[30] E. Vincent, R. Gribonval and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.